MambaPanoptic: A Vision Mamba-based Structured State Space Framework for Panoptic Segmentation

Qing Cheng[1,2] †*, Damiano Bertolini[1,3] †, Wei Zhang[4], Dong Wang[5], Niclas Zeller[6], Daniel Cremers[1,2]

[1] Technical University of Munich, (qing.cheng, cremers)@tum.de
[2] Munich Center for Machine Learning (MCML)
[3] Polytechnic University of Milan, damiano.bertolini@mail.polimi.it
[4] University of Stuttgart, wei.zhang@ifp.uni-stuttgart.de
[5] Wuhan University, timdong@whu.edu.cn
[6] Karlsruhe University of Applied Sciences, niclas.zeller@h-ka.de

Keywords: Panoptic Segmentation, Structured State Space Models, Vision Mamba

Abstract

Panoptic segmentation requires the simultaneous recognition of countable *thing* instances and amorphous *stuff* regions, placing joint demands on long-range context modelling, multi-scale feature representation, and efficient dense prediction. Existing convolutional and transformer-based methods struggle to satisfy all three requirements concurrently: convolutional architectures are limited in their capacity to model long-range dependencies, while transformer-based methods incur quadratic computational cost that is prohibitive at high resolutions. In this paper, we propose MambaPanoptic, a fully Mamba-based panoptic segmentation framework that addresses these limitations through two principal contributions. First, we introduce MambaFPN, a top-down feature pyramid that leverages Mamba blocks to generate globally coherent, multi-scale feature representations with linear computational complexity. Second, we adopt a PanopticFCN-style kernel generator that produces unified *thing* and *stuff* kernels for proposal-free panoptic prediction, enhanced by a QuadMamba-based feature refinement module applied at multiple network stages. Experiments on the Cityscapes and COCO panoptic segmentation benchmarks demonstrate that MambaPanoptic consistently outperforms PanopticDeepLab and PanopticFCN under comparable model sizes, and matches or surpasses Mask2Former on Cityscapes in PQ and AP while requiring fewer parameters.

1. INTRODUCTION

Scene understanding is one of the most important computer vision topics, and panoptic segmentation is one of the most complete and fine-grained tasks in this domain. Panoptic segmentation unifies semantic segmentation and instance segmentation: it estimates a class label for each pixel, and detects the instances and assigns the instance label to them (Kirillov et al., 2019b). This rich and complete scene understanding is so beneficial to various real-world vision-based applications: e.g. autonomous driving, mobile robotics, AR/VR systems, etc, that it has attracted significant attention (Kirillov et al., 2019a, Cheng et al., 2020, Li et al., 2021, Wang et al., 2021, Li et al., 2022, Cheng et al., 2022).

Panoptic segmentation remains challenging because it must simultaneously recognize object instances and delineate amorphous background regions at the pixel level. This requires feature representations that preserve fine local details for accurate boundaries and small objects, while also capturing long-range context to reason about large stuff regions and complex scene layouts. In practice, these competing requirements make panoptic segmentation particularly demanding for high-resolution real-world images.

Conventional methods for panoptic segmentation have depended mostly on Convolutional Neural Networks (CNNs) (Kirillov et al., 2019a) or, more recently, vision transformers (Carion et al., 2020). Convolutional neural networks are effective at capturing local structures through their translation-equivariant inductive bias, but their receptive fields are inherently bounded, limiting their ability to model long-range spatial dependencies. Transformers overcome this limitation through global self-attention, yet do so at a quadratic computational and memory cost with respect to sequence length, which is a significant bottleneck for high-resolution inputs and resource-constrained deployments. Furthermore, the absence of strong inductive biases in transformers typically necessitates large-scale training data and long training time compared to comparably-sized convolutional models.

Structured State Space Models (SSMs), and in particular Mamba architectures, have recently emerged as a promising alternative. Mamba models long-range dependencies through selective state-space dynamics while maintaining linear complexity with respect to the input length (Gu and Dao, 2023). Building on this idea, Vision Mamba extends state-space modeling to visual data by processing image patches with selective bidirectional scanning, enabling both local and global context aggregation. Variants such as the 2D Selective Scan further improve spatial modeling by capturing dependencies along different spatial directions while preserving neighborhood structure (Liu et al., 2025).

In this paper, we propose MambaPanoptic, a fully Mamba-based architecture for panoptic segmentation. Our model integrates a hierarchical Vision Mamba encoder with a novel Mamba Feature Pyramid Network (MambaFPN) — a top-down multi-scale feature pyramid that employs Mamba blocks to produce globally coherent, spatially rich feature representations. Building on this encoder, we adopt a PanopticFCN-style kernel generator

* Corresponding author
† Equal Contribution

(Li et al., 2021) that produces unified *thing* and *stuff* kernels for proposal-free panoptic prediction, and further introduce a QuadMamba-based (Xie et al., 2024) feature refinement module to progressively improve feature quality at multiple network stages. To the best of our knowledge, MambaPanoptic is the first end-to-end Mamba-based architecture designed specifically for panoptic segmentation. SegMAN (Fu et al., 2025) only replaces the ResNet50 (He et al., 2016) with its encoder as the backbone of Mask DINO (Li et al., 2023) for panoptic segmentation. Extensive experiments on the Cityscapes (Cordts et al., 2016) and COCO (Lin et al., 2014) benchmarks demonstrate that MambaPanoptic consistently outperforms CNN-based baselines, including Panoptic-DeepLab (Cheng et al., 2020) and PanopticFCN (Li et al., 2021), and matches or surpasses the transformer-based Mask2Former (Cheng et al., 2022) on Cityscapes in PQ and AP, while requiring fewer parameters.

In sum, our contributions are as follows:

- We present MambaPanoptic, the first fully Mamba-based end-to-end panoptic segmentation architecture with strong performance;
- We propose MambaFPN, a linear-time multi-scale feature pyramid module that jointly provides global context and fine spatial detail for panoptic prediction;
- We also introduce the QuadMamba-based feature refinement module (MFRM) that efficiently enhances feature representations across multiple stages of the network;
- We conduct comprehensive experiments and ablation studies of the proposed method on Cityscapes and COCO panoptic benchmarks, demonstrating competitive performance against representative CNN-based and transformer-based baselines.

2. RELATED WORK

2.1 CNN-based Panoptic Segmentation

Early panoptic segmentation methods were largely built upon CNN-based instance segmentation frameworks, typically by combining an instance branch with a semantic segmentation branch. A representative example is Panoptic FPN (Kirillov et al., 2019a), which extends Mask R-CNN (He et al., 2017) with an additional fully convolutional branch for *stuff* prediction and merges the two outputs through heuristic post-processing. UPSNet (Xiong et al., 2019) further improves this paradigm by adopting a shared deformable-convolution backbone and a lightweight panoptic head with an additional unknown class to better handle conflicts between instance and semantic predictions. Other dual-branch architectures, such as TASCNet (Li et al., 2018) and AUNet (Li et al., 2019), also model *thing* and *stuff* categories separately, but still rely on explicit fusion procedures that may limit the coherence of the final panoptic output.

Another line of work explores bottom-up, proposal-free formulations. Panoptic-DeepLab (Cheng et al., 2020) predicts dense semantic maps together with instance-aware cues, including object center heatmaps and per-pixel offset vectors. This design avoids region proposals and clustering over object candidates while preserving strong spatial consistency in the predicted masks.

More recently, unified single-head CNN architectures have been proposed to further simplify the pipeline. PanopticFCN (Li et al., 2021), for example, formulates both *thing* instances and *stuff* regions as kernel-based predictions, enabling direct mask generation from dense feature maps without relying on bounding boxes or proposal generation. Similarly, the Category-Instance Embedding approach (Gao et al., 2020) learns per-pixel embeddings that jointly encode semantic and instance information within a unified representation space.

Overall, CNN-based methods remain attractive due to their strong locality bias, efficiency, and relatively simple dense prediction pipelines. However, their limited ability to model long-range dependencies can be a disadvantage for panoptic segmentation, where both large-scale scene context and fine-grained spatial details are crucial.

2.2 Transformer-based Panoptic Segmentation

Transformer-based methods have recently become a dominant paradigm for panoptic segmentation by reformulating the task as mask classification. Instead of predicting dense outputs through separate semantic and instance branches, these methods learn a set of queries, each associated with a class label and a segmentation mask. MaskFormer (Cheng et al., 2021) is a representative framework in this direction, showing that panoptic segmentation can be unified through a fixed set of mask queries refined with attention-based decoding. This formulation eliminates the need for handcrafted fusion between semantic and instance predictions and provides a clean end-to-end training framework.

Building on this idea, Mask2Former (Cheng et al., 2022) introduces masked attention, which restricts each query to attend primarily to its predicted spatial support. Combined with a strong multi-scale pixel decoder based on deformable attention, this design improves both efficiency and accuracy, and has become a strong baseline across semantic, instance, and panoptic segmentation tasks. Other transformer-based approaches pursue similar goals with different architectural choices. MaX-DeepLab (Wang et al., 2021) integrates a mask transformer with a CNN backbone and jointly learns pixel features and mask embeddings in a unified framework. Panoptic SegFormer (Li et al., 2022) follows a related query-based design, while introducing separate queries for *thing* and *stuff* categories as well as deep supervision to stabilize optimization.

Compared with CNN-based models, transformer-based methods offer stronger global reasoning and a more unified prediction framework. However, these benefits typically come at the cost of higher computational and memory complexity, particularly for high-resolution dense prediction. This trade-off motivates the investigation of alternative architectures that can retain strong long-range modeling while remaining computationally efficient.

2.3 Vision Mamba-based Segmentation

State-space models (SSMs) led by Mamba (Gu and Dao, 2023) have emerged as efficient alternatives to attention for long-range context modeling with linear-time scaling via input-dependent selective state updates. The 2D Selective Scan of VMamba (Liu et al., 2024) and the bi-directional SSM of Vision Mamba (Liu et al., 2025) bridge 1D sequence modeling to 2D images and demonstrate strong performance on vision tasks, e.g., image classification, object detection, and semantic segmentation. The follow-ups, e.g., MambaVision (Hatamizadeh and Kautz, 2025), GroupMamba (Shaker et al., 2025), MobileMamba (He et al., 2025), SegMAN (Fu et al., 2025), focus on improving the

performance of semantic and instance segmentation. For aerial image segmentation, Mamba-style state-space models have demonstrated notable progress: e.g., dual-branch RS3Mamba (Ma et al., 2024b), encoder–decoder Samba (Ren et al., 2024), large-VHR RS-Mamba (Zhao et al., 2024), and lightweight UNetMamba (Zhu et al., 2024). Meanwhile, the medical variants, e.g. U-Mamba (Ma et al., 2024a), VM-UNet (Ruan et al., 2025), and SegMamba (Xing et al., 2024), also validate the effectiveness of Mamba in encoder–decoder segmentation designs, reinforcing its generality for dense prediction tasks. Collectively, these studies position Vision Mamba backbones as promising for high-resolution semantic segmentation.

Motivated by these advances, we propose a Mamba-based architecture for panoptic segmentation. Our method aims to combine the efficiency and long-range modeling capability of state space models with a unified panoptic prediction framework, providing an alternative to both CNN-based and transformer-based designs.

3. METHOD

In this section, we describe the proposed Mamba-based architecture for panoptic segmentation.

Panoptic segmentation assigns each pixel p a semantic label c, $c \in C$, and a unique instance identity i for each object of thing categories. The label space C is divided into thing classes, which correspond to countable object instances, and stuff classes, which correspond to amorphous background regions.

As illustrated in Figure 1, the proposed method is a unified, proposal-free framework for joint *thing* and *stuff* segmentation, comprising two principal components: a Mamba-based multi-scale feature encoder and a panoptic segmentation head. A Mamba feature refinement module is additionally introduced at multiple stages to improve feature representability. Each component is described in detail below.

3.1 Mamba Multi-Scale Feature Encoder

Panoptic segmentation requires feature representations that preserve fine local structures while also modeling long-range contextual dependencies. To overcome the locality bias of CNNs and the quadratic computation requirement of transformers, we devise a Mamba-based image encoder that aggregates global context with linear complexity with respect to sequence length (Liu et al., 2024). The proposed Mamba-based multi-scale feature encoder consists of a hierarchical backbone and a top-down feature pyramid, as shown in Figure 2.

3.1.1 Mamba backbone We use the SegMAN encoder (Fu et al., 2025) as the backbone, which is a hierarchical hybrid network designed to jointly capture global context and fine-grained spatial detail with linear computational complexity. The SegMAN encoder consists of four blocks. Each block starts with a strided 3×3 convolution for spatial downsampling, followed by a series of Local Attention and State Space (LASS) blocks for feature processing. Each LASS block is structured as a sequential combination of Neighbourhood Attention and the 2D Selective Scan (SS2D) mechanism (Liu et al., 2024). Neighbourhood Attention operates within a fixed local window, preserving spatial precision and translational invariance, while SS2D models long-range dependencies by scanning the flattened feature sequence in four orthogonal directions. The outputs of the two sub-modules are combined via a residual connection and a 1×1 convolution, enabling effective cross-scale feature interaction.

3.1.2 Mamba Feature Pyramid Network Panoptic segmentation consists of semantic segmentation for the *stuff* classes and instance segmentation for the object classes. Stuff classes tend to be dense and cover the full image, while object segments tend to be local and varying-scale. These two sub-tasks impose different requirements for the feature property. Thus, the features are required to be globally coherent to cover the large segments, sufficiently high resolution to encode fine structures, efficiently capturing multi-scale information to handle the varying-size objects, and sufficiently rich semantics to enable reliable class prediction (Kirillov et al., 2019b, Kirillov et al., 2019a, Li et al., 2021). To fulfill these requirements, we design a Mamba-based feature pyramid network (MambaFPN) to generate multi-scale image features for the panoptic head, inspired by FPN (Lin et al., 2017a).

The proposed MambaFPN consists of 4 blocks, each of which fuses the features from two different stages and applies local attention with an SS2D block. The four consecutive blocks of the SegMAN encoder generate four feature maps $\{F_1, F_2, F_3, F_4\}$ at resolutions 1/4, 1/8, 1/16, and 1/32 of the input image, respectively. Starting from the deepest feature map F_4, an SS2D block is applied to produce F_4'. Each subsequent MambaFPN block then upsamples the feature map F_i and integrates the F_{i+1}' by 1×1 convolution and channel-wise addition, and then refines the fused feature map by an SS2D block. This process yields four context-enriched feature maps $\{F_1', F_2', F_3', F_4'\}$ at scales of 1/4, 1/8, 1/16, and 1/32, which are subsequently passed to the panoptic head for panoptic segmentation. The MambaFPN is visualised in Figure 2.

3.2 Panoptic Head

The recent advancements in panoptic segmentation favour the kernel-based (Li et al., 2021), or query-based (Cheng et al., 2021) panoptic head for their effectiveness and simplicity due to their proposal-free and unified single-pass prediction (Li and Chen, 2022). In our work, we adopt the kernel-based panoptic head design to avoid the heavy transformer decoder and self-attention in the query-based panoptic head.

Our panoptic head is inspired by PanopticFCN (Li et al., 2021), which proposes a unified architecture for managing both *thing* and *stuff* prediction by representing them as kernels. We enhance this design with the multi-scale contextual representations produced by MambaFPN and introduce Mamba-based feature refinement at key stages of the head. The panoptic head consists of two branches: a kernel generator and a high-resolution feature encoder. The outputs of both branches are convolved to produce the final panoptic segmentation.

3.2.1 Kernel Generator The kernel generator utilises instance centres and semantic regions to represent *thing* and *stuff* with kernels, respectively. Each kernel can be viewed as a mask with the same resolution as the input feature map. Thus, *thing* and *stuff* have a unified kernel representation. The kernel generator has two sub-modules: a position head for kernel localization and categorization, and a kernel head for producing kernel weights. The position head first refines the input MambaFPN features and predicts heatmaps M^t for *thing* centres, where the peak of a heatmap represents the *thing* center, and score maps M^s for *stuff* regions, where the high response positions of a

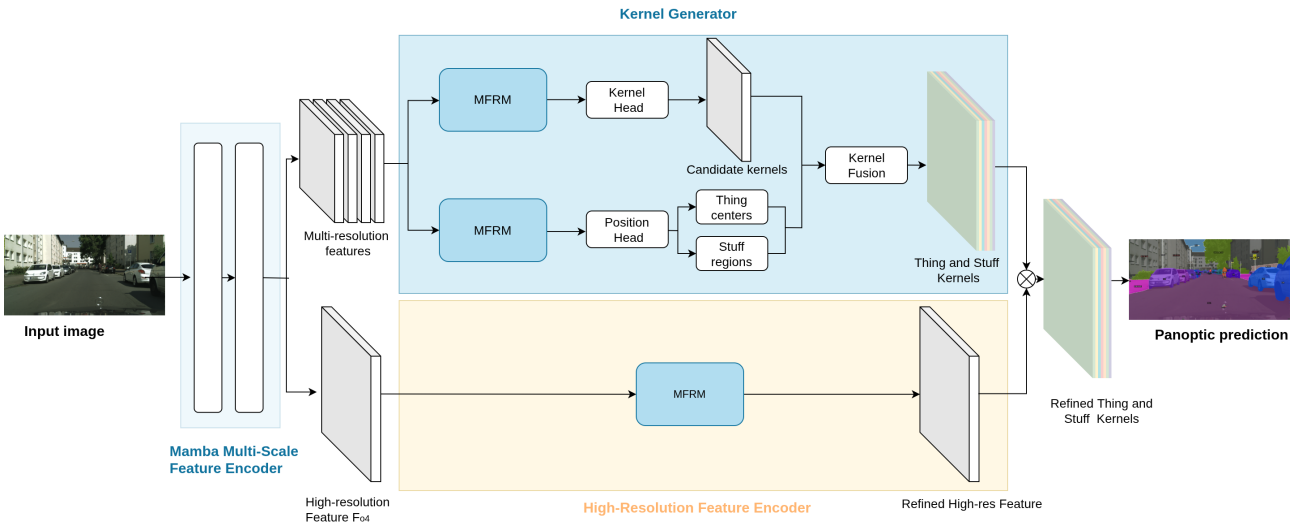

Figure 1. The architecture of the proposed Mamba-based panoptic segmentation network. The MambaFPN takes an image as input the outputs a set of multi-scale feature maps. The kernel generator processes each features and outputs the kernel masks and kernel features at each resolution. Kernel fusion module fuses the kernels from different resolutions into distinct *thing* and *stuff* kernels. The high-resolution feature is refined by the high-resolution feature encoder and then is injected to the kernel features. Finally the panoptic segmentation is estimated based on these final kernels.

score map indicate a *stuff* region. The heatmaps and score maps contribute to the overall kernel masks M. The kernel head is designed to generate the features. It first refines the input MambaFPN features and concatenates the relative coordinates to the refined feature map, and applies a lightweight CNN to generate spatially aware kernel features F. These features are then indexed by the kernel masks to aggregate the features for each kernel, resulting in a set of *thing* kernel features F^t and a set of *stuff* kernel features F^s.

3.2.2 Kernel Fusion The kernel generator works on the four feature maps from the MambaFPN individually, so there can be multiple kernel estimations at different resolutions for the same target. Therefore, we apply a kernel fusion module to integrate the kernels at the different resolutions. We fuse the *thing* and *stuff* kernels differently due to their representations. For *things*, we group kernels across resolutions according to their kernel features: two kernels are identified as the same instance if their cosine similarity is above a given threshold. For *stuff*, we utilize the predicted semantic class and group kernels sharing the same class prediction. The corresponding kernel features are then aggregated via average pooling over the matched kernels. Finally, this yields a unique set of kernel features $\hat{K}$ for each input image, consisting of *thing* kernel features $\hat{K}^t$ and *stuff* kernel features $\hat{K}^s$.

3.2.3 High-Resolution Feature Encoder To better preserve the fine structure and boundaries of the *thing* and *stuff* representation, following PanopticFCN (Li et al., 2021), we utilize the highest-resolution feature map from our MambaFPN to inject high-resolution context into the kernels. Specifically, we first refine the high-resolution feature map, F_1', with our proposed Mamba-based refinement module and use the fused kernel masks M to retrieve the corresponding features and then convolve with the corresponding kernel features $\hat{K}$ to generate the fused kernel features K.

3.2.4 Mamba feature refinement module To refine the features in both the kernel generator and the high-resolution feature encoder, we introduce a Mamba feature refinement module (MFRM) to improve local spatial structure and global contextual awareness, which is built upon a single QuadMamba block (Xie et al., 2024). QuadMamba (Xie et al., 2024) uses a learnable quadtree-based scanning strategy that adaptively partitions the feature map into regions of different granularity. A partition predictor estimates the locality score of each token and determines whether the corresponding region should be further subdivided, enabling coarse-to-fine adaptive processing. In addition, an omnidirectional window-shifting mechanism promotes feature interaction across region boundaries. Compared with fixed-window partitioning (Huang et al., 2024), this design is better suited to objects with varying scales and irregular shapes. We use this module as a lightweight refinement block to enhance feature representations before kernel prediction and final mask decoding.

3.2.5 Panoptic prediction The fused kernel representations are used for the final panoptic prediction. For the *thing* objects, the model predicts a binary mask and a class label over the *thing* labels, and for *stuff* kernels, it estimates per-pixel segmentation over *stuff* classes. Finally, the *thing* and *stuff* masks are merged into a single panoptic map using the heuristic fusion procedure of PanopticFPN (Kirillov et al., 2019a).

3.3 Training

The panoptic prediction task involves two objectives: kernel mask detection and semantic label classification. Accordingly, the model is trained with two complementary loss terms. For the kernel score maps, Focal loss (Lin et al., 2017b) is used for optimization. As the *thing* objects are represented by their centers, we construct the *thing* ground truth as a continuous heatmap generated by placing 2D Gaussian kernels at the center of each object instance. The *stuff* regions for semantic segmentation are formulated commonly with one-hot encoding and bilinearly interpolated to match the resolution of the prediction. The kernel loss is defined as:

$$L_k = \frac{1}{N_t}\sum_i FL(K_i^t, Y_i^t) + \frac{1}{N_s}\sum_i FL(K_i^s, Y_i^s) \quad (1)$$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-225-2026 | © Author(s) 2026. CC BY 4.0 License.

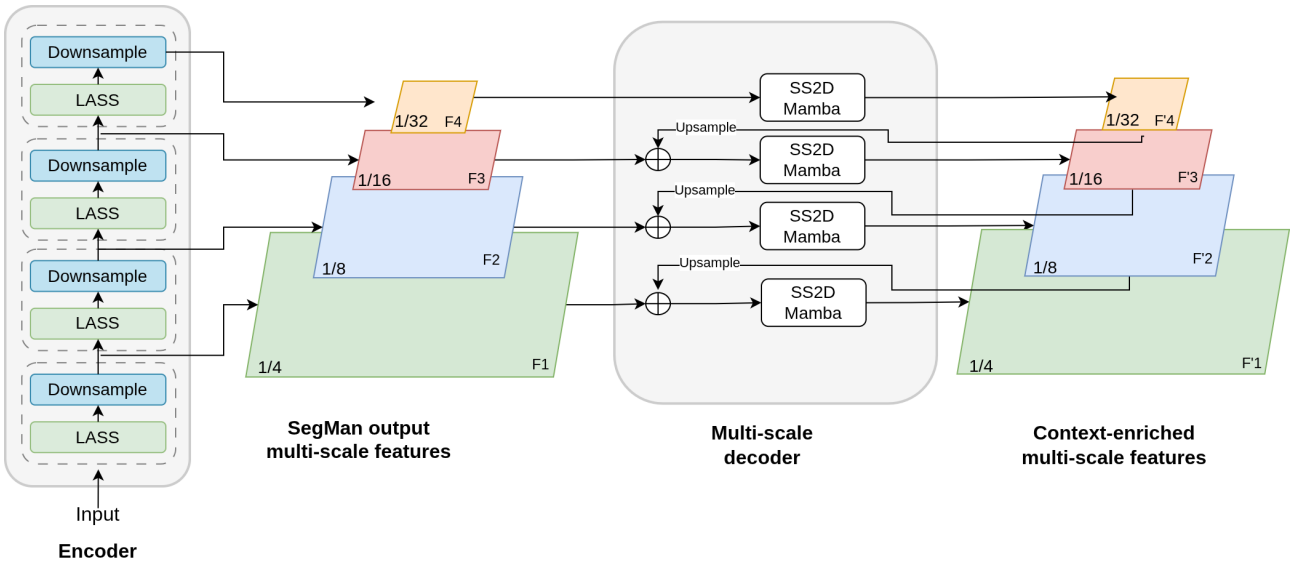

Figure 2. The architecture of the proposed Mamba-based multi-scale feature encoder. The SegMan encoder processes the input image and multi-scale features are extracted. These features are further upsampled to fuse the higher feature map and then processed by the SS2D blocks to generate the context-enriched multi-scale features.

where N_t is the number of *thing* object kernels and N_s is the number of pixels; K_i is the predicted kernel mask and Y_i is the ground truth mask.

For the mask prediction branch, the final mask probabilities P_i are obtained by taking the dot product between the fused kernel features and the refined high-resolution feature map F_1', followed by a sigmoid activation for *thing* masks and softmax for *stuff* regions. Mask prediction is supervised using Dice loss (Milletari et al., 2016) in order to handle the imbalanced label distribution inherent in dense segmentation:

$$L_{seg} = \frac{1}{N} \sum_{i} Dice(P_i, Y_i) \tag{2}$$

The total training loss is a weighted combination of the two terms:

$$L_{total} = \lambda_k L_k + \lambda_{seg} L_{seg} \tag{3}$$

4. EXPERIMENTAL RESULTS

In this section, we present the experiments on two datasets: Cityscapes (Cordts et al., 2016) and COCO 2017 (Lin et al., 2014) for panoptic segmentation quantitatively and qualitatively. We compare against two CNN-based methods, PanopticFCN (Li et al., 2021) and Panoptic-DeepLab (Cheng et al., 2020), and a transformer-based Mask2Former (Cheng et al., 2022). We further conduct ablation studies to assess the contribution of the proposed Mamba-based components.

4.1 Datasets

Experiments are conducted on two widely adopted benchmarks for panoptic segmentation, specifically Cityscapes (Cordts et al., 2016) and COCO 2017 (Lin et al., 2014).

The Cityscapes dataset is designed for urban scene understanding in the context of autonomous driving. Images were collected across 50 cities and are provided at a resolution of 2048×1024 pixels. The dataset comprises 5,000 finely annotated images with pixel-level panoptic labels: 2,975 for training, 500 for validation, and 1,525 for testing. It covers 19 semantic categories, of which 8 are *thing* classes with unique instance labels and 11 are *stuff* classes.

The COCO 2017 dataset (Lin et al., 2014b) is a large-scale benchmark for object detection, instance segmentation, and panoptic segmentation. It contains 118,000 training images, 5,000 validation images, and 20,000 test images, with dense panoptic annotations spanning 133 semantic categories: 80 *thing* classes and 53 *stuff* classes.

4.2 Implementation

We implement our method with Detectron2 (Wu et al., 2019). The base learning rate is set to 0.01, with gradient clipping applied to prevent gradient explosion. Network optimisation is performed using SGD with a weight decay of 1e-4, momentum of 0.9, and a batch size of 24, for a total of 90,000 iterations. The model is evaluated on the validation set every 5,000 iterations, and the checkpoint achieving the highest Panoptic Quality is selected for final evaluation, effectively serving as an early stopping criterion. Random cropping to 512×1024 pixels and CUDA Automatic Mixed Precision (AMP) are applied during training to reduce memory requirements.

4.3 Metrics

To evaluate the performance of panoptic segmentation, our main metric is the Panoptic Quality (PQ) (Kirillov et al., 2019b), a standard metric for the panoptic segmentation task, which jointly captures detection and segmentation performance. It is defined per image as:

$$\text{PQ} = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{4}$$

where TP is the set of true positive matches between predicted segments p and ground-truth segments g with $IoU > 0.5$, FP is the false positives (predicted segments not matched to any

Figure 3. Examples of panoptic predictions on (a) Cityscapes validation set and (b) COCO validation. Each row has two examples. For each example, the prediction overlaid on the image is on the left and the ground truth is on the right.

ground truth), FN is the false negatives (ground truth segments not matched to any prediction) and IoU is the Intersection over Union between the matched prediction and ground truth.

Furthermore, we also report mean Intersection over Union (mIoU) for semantic segmentation and Average Precision (AP) for instance segmentation. These are defined as follows:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{5}$$

where TP_i, FP_i, and FN_i are respectively the true positives, false positives, and false negatives for class i, and N is the number of classes.

$$\text{AP} = \int_0^1 p(r)\,dr \tag{6}$$

Average Precision (AP) is computed as the area under the precision-recall curve, where $p(r)$ is the precision at recall r.

4.4 Results

Table 1. Evaluation results of Panoptic-DeepLab, PanopticFCN, Mask2Former and ours on Cityscapes val set

Method	Backbone	PQ	PQ^{th}	PQ^{st}	AP	mIoU	Param
Panoptic-DeepLab	ResNet50	60.38	50.95	67.24	31.44	77.49	30.3M
PanopticFCN	ResNet50	59.6	52.1	65.1	32.2	76.8	36.6M
Mask2Former	Swin-T	63.9	56.2	67.8	39.1	**80.5**	47.4M
Ours	SegMan	**64.89**	**58.26**	**69.70**	**39.91**	80.18	35.7M

To ensure a fair comparison, all baselines and the proposed model are trained under identical conditions, including the same training schedules, data augmentation strategies, and input resolutions. Backbone networks are selected to maintain comparable parameter counts without modifying the original baseline architectures.

Table 2. Evaluation results of Panoptic-DeepLab, PanopticFCN, Mask2Former and ours on COCO val set

Method	Backbone	PQ	SQ	RQ	AP	mIoU	Param
Panoptic-DeepLab	ResNet50	35.5	77.3	44.7	19.7	40.1	30.3M
PanopticFCN	ResNet50	41.0	81.0	49.6	30.7	43.6	36.6M
Mask2Former	Swin-T	**53.2**	**82.2**	**64.3**	**43.3**	**63.2**	47.4M
Ours	SegMan	45.6	81.5	54.7	34.2	46.4	35.7M

In Table 1, we present the quantitative evaluation results on the Cityscapes validation set. MambaPanoptic outperforms both CNN-based baselines across all reported metrics, and achieves higher PQ and AP than Mask2Former with a slightly lower mIoU, while using 11.7M fewer parameters than the transformer-based model. These results confirm that Mamba-based architectures offer a strong and efficient alternative to both convolutional and transformer designs for panoptic segmentation on urban scene benchmarks.

In Table 2, we present the quantitative evaluation results on the COCO validation set. MambaPanoptic consistently outperforms both CNN-based baselines but shows a performance gap relative to Mask2Former. This gap can be attributed in part to the substantially greater category complexity of COCO. COCO has 80 *thing* classes and 53 *stuff* classes while Cityscapes only has 8 *thing* classes and 11 *stuff* classes. Mask2Former (Cheng et al., 2022) employs 100 learnable queries for COCO panoptic segmentation, each of which can specialise to a distinct category after training, affording high representational capacity

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-225-2026 | © Author(s) 2026. CC BY 4.0 License.

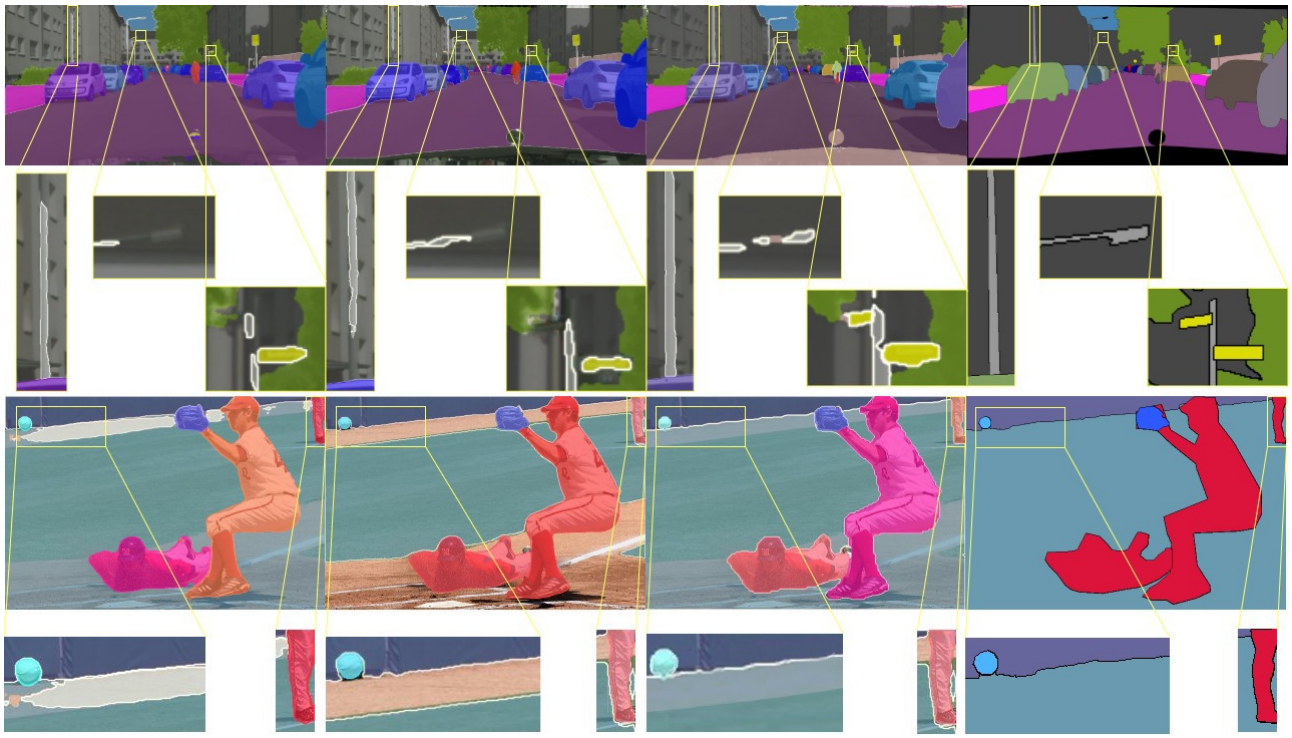

Figure 4. Comparison of CNN-, transformer- and Mamba-based architectures. From left to right: Panoptic-DeepLab (ResNet-50), Mask2Former(Swin-T), our model and finally the ground truth. The two examples are from Cityscapes and COCO, respectively.

for large-vocabulary scenes. In contrast, both MambaPanoptic and PanopticFCN (Li et al., 2021) generate kernels dynamically from each input image, without the benefit of persistent, category-specific query representations. This architectural distinction means that query-based methods carry a representational advantage in high-category settings, at the cost of heavier decoders and quadratic self-attention.

In Figure 3, we present qualitative panoptic segmentation results on both validation sets. The proposed model correctly segments both *thing* instances and *stuff* regions, including objects that are distant or partially occluded. Predictions near object boundaries occasionally exhibit imprecision. Figure 4 provides a direct qualitative comparison of small and distant structures in both datasets among Panoptic-DeepLab, Mask2Former, and MambaPanoptic, which shows the effectiveness of our proposed method compared to Panoptic-DeepLab and Mask2Former.

4.5 Ablations

Table 3. Ablation results of Mamba-based modules on the Cityscapes val set.

Method	PQ	PQth	PQst	AP	mIoU
w/o Mamba Encoder	59.85	52.43	64.49	32.54	77.01
w/o MambaFPN	62.49	53.42	66.09	34.82	78.14
w/o QuadMamba	64.39	58.03	69.28	39.13	79.95
Ours	**64.89**	**58.26**	**69.70**	**39.91**	**80.18**

In this ablation section, we mainly verify the effectiveness of the three Mamba-based modules introduced. For each ablation, we use the same training and evaluation settings except for the replacement of the test module. In Table 3, we show the ablation results on the Cityscapes dataset. *Ours* represents the proposed model; *w/o Mamba Encoder* represents the model that uses the ResNet50 as the backbone instead of the SegMan encoder; *w/o MambaFPN* represents the model that uses CNN FPN instead of the MambaFPN for multi-scale feature generation; *w/o QuadMamba* represents the model that uses a CNN instead of the QuadMamba for feature refinement. The results in Table 3 confirm that each proposed Mamba-based module contributes positively and independently to the overall panoptic segmentation performance.

5. CONCLUSION

In this paper, we have presented MambaPanoptic, a Mamba-based framework for panoptic segmentation. The proposed architecture combines a multi-scale feature encoder based on the SegMAN backbone and the proposed MambaFPN, together with a PanopticFCN-style kernel-based panoptic head. To further enhance feature quality, we incorporated a QuadMamba-based refinement module into both kernel generation and high-resolution feature refinement. Experimental results on the Cityscapes and COCO panoptic benchmarks demonstrate that the proposed method is effective and competitive, showing clear improvements over strong CNN-based baselines and strong performance on Cityscapes with fewer parameters than Mask2Former. These results suggest that state-space models are a promising alternative to conventional convolutional and transformer-based architectures for panoptic segmentation.

Despite these encouraging results, the proposed method still has several limitations. First, as observed in the qualitative results, the model occasionally struggles near object boundaries and thin structures. We hypothesize that this behaviour is attributable to a trade-off inherent in Mamba-based scanning mechanisms, which are effective at modeling long-range dependencies but may be less suited to capturing high-frequency local details. Incorporating boundary-aware supervision or more specialized high-resolution refinement may help alleviate this issue. Second, although the kernel-based head is efficient, the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-225-2026 | © Author(s) 2026. CC BY 4.0 License.

method still lags behind stronger query-based transformer models on highly complex datasets such as COCO, which contain many semantic categories and crowded scene layouts. This suggests that query-based decoding may provide higher representation capacity in such settings. Exploring a hybrid design that combines efficient Mamba-based feature modeling with a lightweight query-based decoder is therefore a promising direction for future work.

In future work, we plan to investigate larger Mamba-based backbones, improve prediction quality near boundaries and thin structures, explore more tightly integrated Mamba-based pixel decoders and panoptic heads, and extend the framework to related dense prediction tasks such as aerial-image and 3D panoptic segmentation, where efficient long-range context modeling may offer even greater benefits.

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. *European conference on computer vision*, Springer, 213–229.

Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., Chen, L.-C., 2020. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12475–12485.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12037–12047.

Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34, 17864–17875.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Fu, Y., Lou, M., Yu, Y., 2025. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. arXiv:2412.11890.

Gao, N., Shan, Y., Zhao, X., Huang, K., 2020. Learning category-and instance-aware pixel embedding for fast panoptic segmentation. *European conference on computer vision*, Springer, 411–427.

Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hatamizadeh, A., Kautz, J., 2025. Mambavision: A hybrid mamba-transformer vision backbone. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25261–25270.

He, H., Zhang, J., Cai, Y., Chen, H., Hu, X., Gan, Z., Wang, Y., Wang, C., Wu, Y., Xie, L., 2025. Mobilemamba: Lightweight multi-receptive visual mamba network. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4497–4507.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C., 2024. LocalMamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*.

Kirillov, A., Girshick, R., He, K., Dollár, P., 2019a. Panoptic feature pyramid networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6399–6408.

Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P., 2019b. Panoptic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9404–9413.

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L. M., Shum, H.-Y., 2023. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3041–3050.

Li, J., Raventos, A., Bhargava, A., Tagawa, T., Gaidon, A., 2018. Learning to fuse things and stuff. *arXiv preprint arXiv:1812.01192*.

Li, X., Chen, D., 2022. A survey on deep learning-based panoptic segmentation. *Digital Signal Processing*, 120, 103283.

Li, Y., Chen, X., Zhu, Z., Xie, L., Huang, G., Du, D., Wang, X., 2019. Attention-guided unified network for panoptic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7026–7035.

Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., Jia, J., 2021. Fully convolutional networks for panoptic segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14207–14216.

Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., Lu, T., 2022. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8734–8743.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *European conference on computer vision*, Springer, 740–755.

Liu, X., Zhang, C., Huang, F., Xia, S., Wang, G., Zhang, L., 2025. Vision mamba: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y., 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37, 103031–103063.

Ma, J., Li, F., Wang, B., 2024a. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*.

Ma, X., Zhang, X., Pun, M.-O., 2024b. Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, Ieee, 565–571.

Ren, L., Liu, Y., Lu, Y., Shen, Y., Liang, C., Chen, W., 2024. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *arXiv preprint arXiv:2406.07522*.

Ruan, J., Li, J., Xiang, S., 2025. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *ACM Trans. Multimedia Comput. Commun. Appl.* https://doi.org/10.1145/3767748.

Shaker, A., Wasim, S. T., Khan, S., Gall, J., Khan, F. S., 2025. Groupmamba: Efficient group-based visual state space model. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14912–14922.

Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C., 2021. Max-deeplab: End-to-end panoptic segmentation with mask transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5463–5474.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2. `https://github.com/facebookresearch/detectron2`.

Xie, F., Zhang, W., Wang, Z., Ma, C., 2024. Quadmamba: Learning quadtree-based selective scan for visual state space model. *Advances in Neural Information Processing Systems*, 37, 117682–117707.

Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L., 2024. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation . *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, LNCS 15008, Springer Nature Switzerland.

Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R., 2019. Upsnet: A unified panoptic segmentation network. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8818–8826.

Zhao, S., Chen, H., Zhang, X., Xiao, P., Bai, L., Ouyang, W., 2024. Rs-mamba for large remote sensing image dense prediction. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhu, E., Chen, Z., Wang, D., Shi, H., Liu, X., Wang, L., 2024. Unetmamba: An efficient unet-like mamba for semantic segmentation of high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*.